

REMARKS

The Office Action mailed September 29, 2009 (*hereinafter* Action) has been received and its contents carefully considered. The Applicants thank the Examiner for the careful consideration of this application. No claims have been amended or added. Claims 2-7, 9-23, 25, 45, 49, 52-53, and 60-61 were previously canceled. Claims 1, 46, 62, and 63 are the independent claims.

Accordingly, upon entry of this Amendment, claims 1, 8, 24, 26-44, 46-48, 50-51, 54-59, and 62-66 are pending in the application with claims 54-59 withdrawn from consideration. Based on the following remarks, the Applicants respectfully request that the Examiner reconsider all outstanding rejections, and that they be withdrawn. Reconsideration is respectfully requested.

Claim Rejections – 35 U.S.C. § 102

Beginning on page 3, the Action has rejected claims 1, 8, 24, 26-34, 36, 38-44, 46-48, 50-51, and 62-66 35 U.S.C. § 102(b) as being anticipated by Diligenti et al., FOCUSED CRAWLING USING CONTEXT GRAPHS, 26th International Conference on Very Large Databases, VLDB 2000, pages 527-534 (*hereinafter* Diligenti). The Applicants respectfully traverse this rejection.

The crawling method described by Diligenti is largely based on the work of Soumen Chakrabarti, Martin van den Berg and Byron Dom: “Focused crawling: a new approach to topic-specific Web resource discovery” (*hereinafter* Chakrabarti). As discussed in the Response filed July 13, 2009, the method of Chakrabarti differs from an embodiment of the current claimed invention.

Chakrabarti

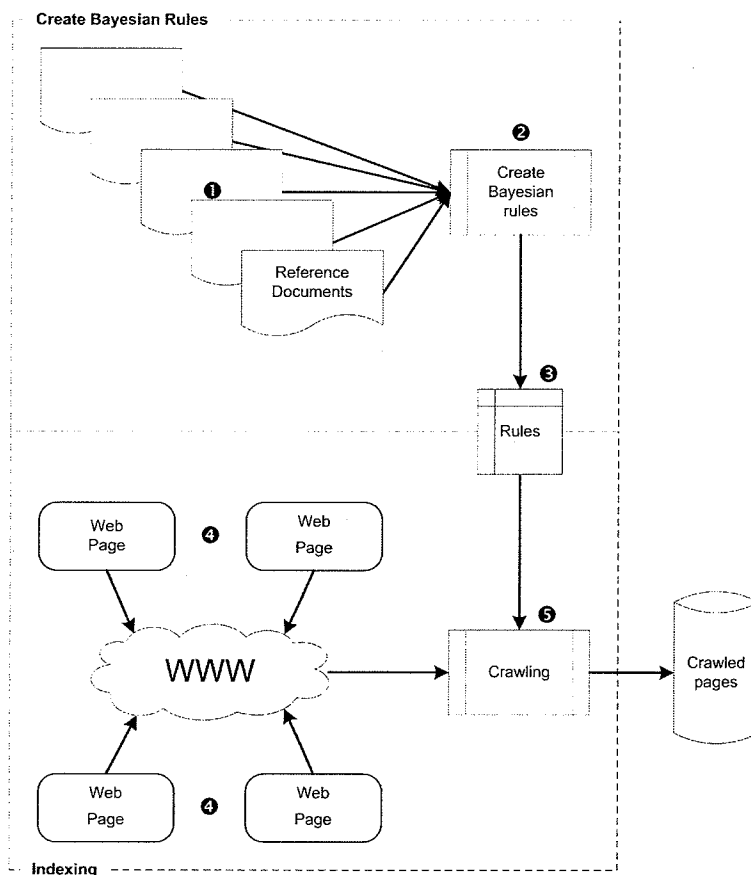


Figure 1: The Method of Chakrabarti

Chakrabarti discloses a method where crawling is assisted by a statistical comparison to a set of “reference” documents, figure 1, ❶, using Bayesian rules. This method does not apply fixed rules to select what pages to index, but uses a “fuzzy” set of probabilistic rules, figure 1, ❸, that are established by performing statistics, figure 1, ❷, on a number of reference documents, figure 1, ❶. In short, Chakrabarti selects a number of reference documents, figure 1, ❶, that they believe are relevant for the focus they want, performs some statistics, figure 1, ❷, on those documents, using the calculated Bayes rules, figure 1, ❹, to select the order of pages, figure 1, ❹, to crawl, figure 1,

⑤, based on probability. Chakrabarti does not explicitly include or exclude pages based on the calculations but uses the calculations to crawl “relevant” pages faster than by traversing the net in a more traditional breadth first manner.

The method Chakrabarti describes should be considered a personal document retrieval system for web pages more than a search engine as it is meant to crawl and display pages in real time to a user. The Chakrabarti system is thus intended to be a desktop application employed on the user's personal computer rather than a web-based search engine.

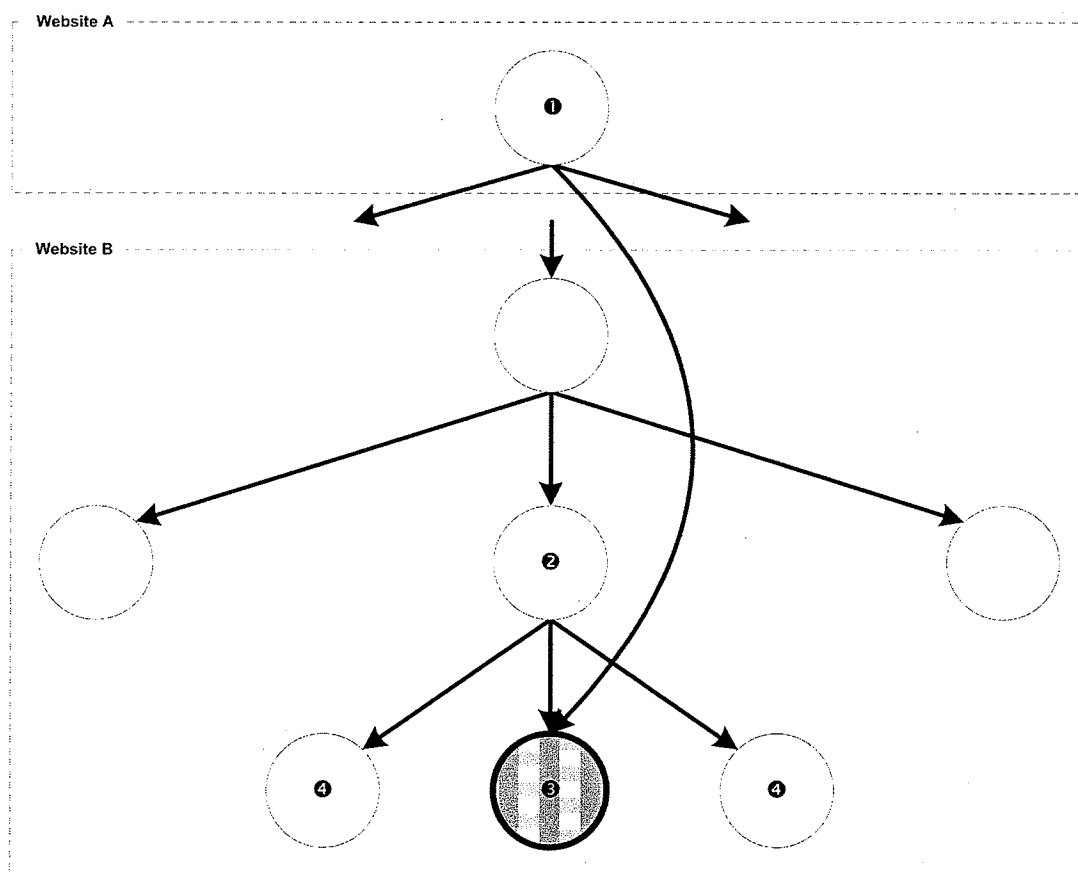


Figure 2: How Chakrabarti Crawls Pages

Diligenti

As shown in figures 2 and 3, the method of Diligenti refines the method of Chakrabarti by refining the way links are followed to get to relevant pages quicker by assuming that web pages on the same “level”, ④, as the current, relevant page, ③, probably are relevant too.

Chakrabarti follows direct links only, which in figure 2 starts at a relevant page on Website A, ①, that leads to a relevant page on Website B, ③. The possibly relevant pages on the same level, ④, are not found this way as there are no direct links from Website A to a page on a higher level than the relevant page, ③, and thus no links to the possibly relevant pages on the same level, ④.

The situation for Diligenti is similar, as shown in figure 3, except that Diligenti uses an external search engine like Google to find pages, ④, on the same level as the relevant page already found, ③, as indicated with the dotted arrows.

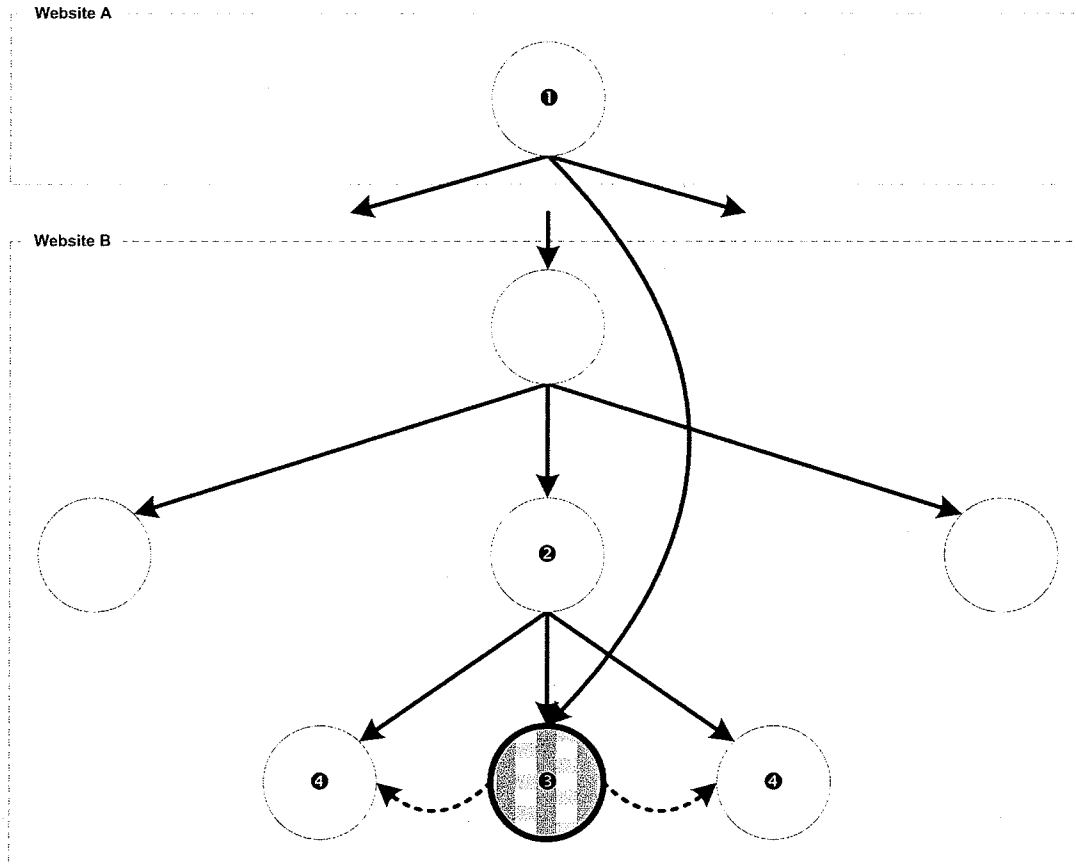


Figure 3: How Diligenti Crawl Pages

The end result, when it comes to the basic search strategy is that Diligenti does not differ from Chakrabarti except for the use of an external search engine to try to find pages to search on the same “level” (meaning with a common page on the same site linking to them, ②).

The method Diligenti describes should be considered a personal document retrieval system for web pages more than a search engine as it is meant to crawl and display pages in real time to a user. For example, section 5, page 7 of Diligenti recites “We have no doubt that further improvement of focused crawling will soon make crawling not only the privilege of large companies that can afford expensive infrastructures but a personal tool that is widely available for retrieving information on the world wide web.” Thus, Diligenti's system is like Chakrabarti's,

intended to be a desktop application employed on a user's personal computer rather than a web-based search engine.

Diligenti describes a system where a user, by locating a number of (reference) documents/pages she finds relevant and then pointing the described system to these, enables a system crawl the web to hopefully quickly find pages that are similar to the reference documents/pages through statistical measures, using Bayesian rules.

Claim 1

The Applicants respectfully traverse this rejection for at least the following reasons.

First, Diligenti does not disclose “filtering, by said search engine, **subject specific content** of each said object visited to determine a relevance of said subject specific content thereof to said predefined particular subject” as recited by claim 1 (emphasis added).

On page 3, the Action aligns the recited features with page 5, section 3.3 and page 4, column 2, paragraph 3 of Diligenti. However, Diligenti discloses using a set of reference documents to “train” (“This representation is used to train a set of classifiers, which are optimized to detect and assign documents to different categories based on the expected link distance from the document to the target document,” Diligenti, page 3, section 3, paragraph 1) a statistically based system to configure Bayesian rules (“we use a modification of the *Naive Bayes Classifier*,” Diligenti, page 4, section 3.2, paragraph 4) that are subsequently used to order pages for the selection of what page to *crawl* next. Diligenti also discloses that “During the crawling phase, new context graphs can periodically be built for every topically relevant element found in queue 0” page 5, section 3.3, last paragraph. Thus, if a page does not match the Bayesian rules, the sub-tree of pages linked to from this page is discarded, which could easily prune relevant pages from the crawling. This was noted

by the Examiner on page 15 of the action, “sites deemed irrelevant (e.g. not meeting the minimum confidence threshold) are categorized as ‘other’ and not crawled further” (emphasis added). In sum, a properly trained Bayesian filter does not make the selection subject specific.

Therefore, Diligenti does not disclose “filtering, by said search engine, subject specific content of each said object visited to determine a relevance of said subject specific content thereof to said predefined particular subject” as recited by claim 1.

Thus, Diligenti does not disclose each and every feature of claim 1 as required by 35 U.S.C. § 102, and claim 1 is patentable for a first reason.

Second, Diligenti does not disclose:

comparing said decomposed components of said objects to said subject specific terminology of the lexicon to determine whether each said object is a subject specific relevant object, wherein said comparing comprises:

(i) assigning a weight to each of said words, terms and expressions comprising the subject specific terminology of the lexicon;

as recited by claim 1.

On page 4, the Action aligns the recited features with page 4-5 of Diligenti. However, on page 4, section 3.2, paragraph 4, Diligenti discloses that “we use a modification of the *Naive Bayes Classifier*.” Bayesian filtering cannot easily be fine tuned, as the only way to change the scores for pages is by changing the set of reference documents used to train the system. If the documents used for training are too few, the Bayesian rules will tend to select a narrow set of pages and thus miss many relevant pages. On the other hand, if too many documents are used to train the system, the Bayesian rules will tend to select a very wide set of pages and thus possibly include too many irrelevant pages. Another difficulty with Bayesian statistics is that it can be very difficult to select a

representative set of reference documents for training as incidental coincidences in the selected documents could lead to unpredictable results.

Furthermore, as described in page 4, section 3.2, Diligenti uses Term Frequency Inverse Document Frequency (TF-IDF) to score pages, where the weight of a phrase in a page depends on the number of times the phrase occurs in the page (the more times the higher the weight) and how many times the phrase occurs in the reference documents (the fewer time the higher the weight). "If two phrases have the same number of occurrences in a document, the TF-IDF value of the less common phrase will be higher." Diligenti, page 4, column 2, paragraph 1. This way of creating weights can easily lead to the selection of irrelevant pages for inclusion as irrelevant phrases that seldom occur in the reference documents but often in the page looked at would weigh this phrase highly, and this way of creating weights is also quantitative and not qualitative, as irrelevant phrases could easily trigger inclusion. Hence, Diligenti does not have a qualitative approach to creating the classifier vector, but rather a quantitative approach where words appearing often in a page but seldom in the reference corpus are deemed relevant.

Thus, Diligenti does not disclose:

comparing said decomposed components of said objects to said subject specific terminology of the lexicon to determine whether each said object is a subject specific relevant object, wherein said comparing comprises:

- (i) assigning a weight to each of said words, terms and expressions comprising the subject specific terminology of the lexicon;

as recited by claim 1.

Therefore, Diligenti does not disclose each and every feature of claim 1 as required by 35 U.S.C. § 102, and claim 1 is patentable for a second reason.

Third, Diligenti does not disclose, “if a said word, term or expression comprising the object matches a corresponding said word, term or expression comprising the subject specific terminology of the lexicon, adding a corresponding weight thereof to a **cumulative total**” as recited by claim 1 (emphasis added).

On page 5, the Action aligns the recited feature with page 5, equation 3 of Diligenti. However, equation 3 of Diligenti calculates a probability of how well all the words in the current page match the words in the reference corpus. “The **probability** that a vector element w_i occurs in documents of class c_j is $P(w_i|c_j)$.” Diligenti, page 5, column 1, paragraph 1 (emphasis added).

Therefore, Diligenti does not disclose, “if a said word, term or expression comprising the object matches a corresponding said word, term or expression comprising the subject specific terminology of the lexicon, adding a corresponding weight thereof to a cumulative total” as recited by claim 1.

Thus, Diligenti does not disclose each and every feature of claim 1 as required by 35 U.S.C. § 102, and claim 1 is patentable for a third reason.

Fourth, Diligenti does not disclose “determining any of said objects to be a subject specific relevant object if the cumulative total surpasses a predefined threshold value” as recited by claim 1.

On page 5, the Action aligns the recited feature with page 5, section 3.3 of Diligenti. However, section 3.3 of Diligenti discloses a way of sorting pages to select the next page for crawling. “When the crawler needs the next document to move to, it pops from the first non-empty queue.” Diligenti, page 5, column 2, paragraph 4.

Thus, Diligenti does not disclose “determining any of said objects to be a subject specific relevant object if the cumulative total surpasses a predefined threshold value” as recited by claim 1.

Therefore, Diligenti does not disclose each and every feature of claim 1 as required by 35 U.S.C. § 102, and claim 1 is patentable for a fourth reason.

For at least the four reasons given above, claim 1 is allowable, and the Applicants respectfully request removal of this rejection.

Claim 46

The Applicants respectfully traverse this rejection for at least the following reason.

Claim 46 contains similar language to allowable claim 1 and is, therefore, allowable for at least the reasons given for claim 1. The Applicants respectfully request removal of this rejection.

Claim 62

The Applicants respectfully traverse this rejection for at least the following reasons.

First, Diligenti does not disclose “filtering, by said search engine, **subject specific content** of each said object visited to determine relevance of said subject specific content thereof to said predefined particular subject” as recited by claim 62 (emphasis added).

On page 15, the Action aligns the recited features with page 5, section 3.3 and page 4, column 2, paragraph 3 of Diligenti. However, Diligenti discloses using a set of reference documents to “train” (“This representation is used to train a set of classifiers, which are optimized to detect and assign documents to different categories based on the expected link distance from the document to the target document,” Diligenti, page 3, section 3, paragraph 1) a statistically based system to configure Bayesian rules (“we use a modification of the *Naive Bayes Classifier*,” Diligenti, page 4, section 3.2, paragraph 4) that are subsequently used to order pages for the selection of what page to *crawl* next. Diligenti also discloses that “During the crawling phase, new context graphs can periodically be built for every topically relevant element found in queue 0” page

5, section 3.3, last paragraph. Thus, if a page does not match the Bayesian rules, the sub-tree of pages linked to from this page is discarded, which could easily prune relevant pages from the crawling. This was noted by the Examiner on page 15 of the action, “sites deemed irrelevant (e.g. not meeting the minimum confidence threshold) are categorized as ‘other’ and not crawled further” (emphasis added). In sum, a properly trained Bayesian filter does not make the selection subject specific.

Therefore, Diligenti does not disclose “filtering, by said search engine, subject specific content of each said object visited to determine relevance of said subject specific content thereof to said predefined particular subject” as recited by claim 62.

Thus, Diligenti does not disclose each and every feature of claim 62 as required by 35 U.S.C. § 102, and claim 62 is patentable for a first reason.

Second, Diligenti does not disclose “**presenting** one or more of said components of each of said objects **to a human editor via a human computer interface**” as recited by claim 62 (emphasis added).

On page 15, the action aligns the recited feature with page 2, column 2, paragraph 4, of Diligenti. However, this section of Diligenti discloses training a crawler based on web sites that have been previously selected for their training potential. On page 2, column 2, paragraph 4, Diligenti discloses that this method is inappropriate, “However, this approach places a burden on the user to specify representative web sites” (emphasis added).

Thus, Diligenti does not disclose “presenting one or more of said components of each of said objects to a human editor via a human computer interface” as recited by claim 62.

Therefore, Diligenti does not disclose each and every feature of claim 62 as required by 35 U.S.C. § 102, and claim 62 is patentable for a second reason.

Third, Diligenti does not disclose “**permitting the human editor to deem a said object to be a subject specific relevant object** if the human editor determines any of said components comprising said object to be within said predefined particular subject” as recited by claim 62 (emphasis added).

On page 16, the action aligns the recited feature with page 2, column 2, paragraph 4, of Diligenti. However, as discussed directly above, this section of Diligenti discloses training a crawler based on web sites that have been previously selected for their training potential. On page 2, column 2, paragraph 4, Diligenti discloses that this method is inappropriate, “However, this approach places a burden on the user to specify representative web sites” (emphasis added).

Thus, Diligenti does not disclose “permitting the human editor to deem a said object to be a subject specific relevant object if the human editor determines any of said components comprising said object to be within said predefined particular subject” as recited by claim 62.

Therefore, Diligenti does not disclose each and every feature of claim 62 as required by 35 U.S.C. § 102, and claim 62 is patentable for a third reason.

Fourth, Diligenti does not disclose “**permitting the human editor to deem a said object to not be a subject specific relevant object** if the human editor determines any of said components comprising said object to not be within said predefined particular subject” as recited by claim 62 (emphasis added).

On page 15, the action aligns the recited feature with page 2, column 2, paragraph 4, of Diligenti. However, this section of Diligenti discloses training a crawler based on web sites that

have been previously selected for their training potential. On page 2, column 2, paragraph 4, Diligenti discloses that this method is inappropriate, “However, this approach places a burden on the user to specify representative web sites” (emphasis added).

Thus, Diligenti does not disclose “permitting the human editor to deem a said object to not be a subject specific relevant object if the human editor determines any of said components comprising said object to not be within said predefined particular subject” as recited by claim 62.

Therefore, Diligenti does not disclose each and every feature of claim 62 as required by 35 U.S.C. § 102, and claim 62 is patentable for a fourth reason.

For at least the four reasons given above, claim 62 is allowable, and the Applicants respectfully request removal of this rejection.

Claim 63

The Applicants respectfully traverse this rejection for at least the following reasons.

First, Diligenti does not disclose “filtering, by said search engine, **subject specific content** of each said object visited to determine relevance of said subject specific content thereof to said predefined particular subject” as recited by claim 63 (emphasis added).

Beginning on page 17, the Action aligns the recited features with page 5, section 3.3 and page 4, column 2, paragraph 3 of Diligenti. However, Diligenti discloses using a set of reference documents to “train” (“This representation is used to train a set of classifiers, which are optimized to detect and assign documents to different categories based on the expected link distance from the document to the target document,” Diligenti, page 3, section 3, paragraph 1) a statistically based system to configure Bayesian rules (“we use a modification of the *Naive Bayes Classifier*,” Diligenti, page 4, section 3.2, paragraph 4) that are subsequently used to order pages for the

selection of what page to *crawl* next. Diligenti also discloses that “During the crawling phase, new context graphs can periodically be built for every topically relevant element found in queue 0” page 5, section 3.3, last paragraph. Thus, if a page does not match the Bayesian rules, the sub-tree of pages linked to from this page is discarded, which could easily prune relevant pages from the crawling. This was noted by the Examiner on page 15 of the action, “sites deemed irrelevant (e.g. not meeting the minimum confidence threshold) are categorized as ‘other’ and not crawled further” (emphasis added). In sum, a properly trained Bayesian filter does not make the selection subject specific.

Therefore, Diligenti does not disclose “filtering, by said search engine, subject specific content of each said object visited to determine relevance of said subject specific content thereof to said predefined particular subject” as recited by claim 63.

Thus, Diligenti does not disclose each and every feature of claim 63 as required by 35 U.S.C. § 102, and claim 63 is patentable for a first reason.

Second, Diligenti does not disclose “comparing said decomposed components of said objects to said subject specific terminology of the lexicon to determine whether each said object is a **subject specific** relevant object, wherein a said object is deemed to be a subject specific relevant object if at least one component thereof matches a corresponding subject specific terminology of the lexicon” as recited by claim 63 (emphasis added).

On page 18, the Action aligns the recited feature with page 4, column 2, paragraph 3 and page 5, section 3.3. However, Diligenti page 3, column 1, paragraph 4, discloses “An initiation phase when a set of context graphs and associated classifiers are constructed for each of the seed

documents.” Hence, Diligenti does not determine whether a web page is relevant in relation to a specific subject but determines how well a page relates to a set of reference documents.

Thus, Diligenti does not disclose “comparing said decomposed components of said objects to said subject specific terminology of the lexicon to determine whether each said object is a subject specific relevant object, wherein a said object is deemed to be a subject specific relevant object if at least one component thereof matches a corresponding subject specific terminology of the lexicon” as recited by claim 63.

Therefore, Diligenti does not disclose each and every feature of claim 63 as required by 35 U.S.C. § 102, and claim 63 is patentable for a second reason.

For at least the two reasons given above, claim 63 is allowable, and the Applicants respectfully request removal of this rejection.

Claims 8, 24, 26-34, 36, 38-44, 47-48, 50-51, and 64-66

Claims 8, 24, 26-34, 36, 38-44, 47-48, 50-51, and 64-66 are either directly or indirectly dependent from allowable claims 1 or 46. Therefore, claims 8, 24, 26-34, 36, 38-44, 47-48, 50-51, and 64-66 are allowable, at least, for being dependent from an allowable claim.

The Applicants respectfully request withdrawal of this rejection.

Claim Rejections – 35 U.S.C. §103

Beginning on page 19 of the Action, the Examiner has rejected claims 35 and 37 under 35 U.S.C. §103(a) as being unpatentable over Diligenti in view of Filippo Menczer et al., EVALUATING TOPIC-DRIVEN WEB CRAWLERS (2001) (*hereinafter* Menczer). The Applicants respectfully traverse this rejection.

Menczer discloses “three novel approaches for assessing and comparing topic driven crawlers.” Menczer, page 1, section 1, last paragraph.

Second, the article “appl[ies] this evaluation framework to compare three types of crawlers.” *Id.* This objective fails, however, as the authors seem to confuse “Crawler” and “Ranking” in as much as they claim to be evaluating the crawlers’ ability to differentiate between different topics they actually evaluate the ranking-mechanisms of the search facilities in question.

As a point of comparison, Menczer specifies a system used to rank documents already harvested by a crawler according to relevance for a specific topic. The system is used to compare different crawler and ranking strategies to see which are best at finding documents relevant for a specific topic. This system does not filter documents according to subject/topic.

The first component of a search engine is the mechanism designed to travel networked information resources — the “crawler.” The crawler is what differentiates search engines from other search facilities. The crawling process of General Purpose Search Engines will consider all information relevant, whereas the crawling process of an embodiment of the present invention, a Subject Specific Search Engine, will not as it will only consider information relating to the specific subject, e.g. Law or Medicine, relevant.

The second component is the Data Storage or the Index, this is where all the information retrieved by the crawler is stored and indexed and made searchable to users (or searchers).

The third component is a ranking mechanism that will attempt rank the results to a searcher’s query string in such a way that the most relevant results are listed first on the results list. This mechanism assesses the relative relevance of results to a query string—the ranking is thus a

function of the query string. The ranking of results is independent and as such unrelated to the function of the crawler.

Thus, Menczer in failing to recognize the different search technologies also fails to recognize that they are not “assessing and comparing topic driven crawlers,” e.g. Google which is a General Purpose Search Engine using PageRank as their ranking mechanism is not a “topic driven crawler” and neither are any of the others they compare.

Finally, the Menczer article fails to recognize the difference between different types of search facilities and between the different component of search engines by labeling all “Crawlers”—what they seem to refer to is any type of search facility irrespective of what kind of technology that produces the results. In essence Menczer operates with a black box (of unknown content technology wise) called the “Crawler,” which covers the harvesting of data as well as the ranking of search results.

What they are in reality testing is the ranking mechanism of a specific search facility. Thus their goal is really to determine which ranking mechanism of the ones in question is the better at listing results relevant to the given topic first on their results lists.

Claims 35 and 37

Claims 35 and 37 are indirectly dependent from allowable claim 1. Menczer, as discussed above, does not remedy the deficiencies in Diligenti. Therefore, claims 35 and 37 are allowable for at least being dependent from an allowable claim.

Accordingly, the Applicants respectfully request removal of this rejection.

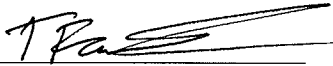
CONCLUSION

All of the stated grounds of rejection have been properly traversed, accommodated, or rendered moot. The Applicants therefore respectfully request that the Examiner reconsider all presently outstanding rejections and that they be withdrawn. The Applicants believe that a full and complete reply has been made to the outstanding Office Action and, as such, the present application is in condition for allowance. If the Examiner believes, for any reason, that personal communication will expedite prosecution of this application, the Examiner is hereby invited to telephone the undersigned at the number provided.

In view of the above amendments, the Applicants believe the pending application is in condition for allowance.

Dated: 10/28/2009

Respectfully submitted,

By 
Todd Richard Farnsworth
Registration No.: 65,432
Cameron H. Tousi
Registration No.: 43,197
VENABLE LLP
P.O. Box 34385
Washington, DC 20043-9998
(202) 344-4000
(202) 344-8300 (Fax)
Attorney/Agent For Applicants

1080822